Disco: Disentangled Control for Realistic Human Dance Generation

Abstract

Generative AI has made significant strides in computer vision, particularly in text-driven image/video synthesis (T2I/T2V). Despite the notable advancements, it remains challenging in human-centric content synthesis such as realistic dance generation. Current methodologies, primarily tailored for human motion transfer, encounter difficulties when confronted with real-world dance scenarios (e.g., social media dance), which require to generalize across a wide spectrum of poses and intricate human details. In this paper, we depart from the traditional paradigm of human motion transfer and emphasize two additional critical attributes for the synthesis of human dance content in social media contexts: (i) Generalizability: the model should be able to generalize beyond generic human viewpoints as well as unseen human subjects, backgrounds, and poses; (ii) Compositionality: it should allow for the seamless composition of seen/unseen subjects, backgrounds, and poses from different sources. To address these challenges, we introduce Disco, which includes a novel model architecture with disentangled control to improve the compositionality of dance synthesis, and an effective human attribute pre-training for better generalizability to unseen humans. Extensive qualitative and quantitative results demonstrate that DISCO can generate high-quality human dance images and videos with diverse appearances and flexible motions. Code is available at https://disco-dance.github.io/.